

CRACKLE: THE PERSISTENT HOMOLOGY OF NOISE

BY ROBERT J. ADLER OMER BOBROWSKI AND SHMUEL WEINBERGER

Electrical Engineering, Technion – Israel Institute of Technology

Department of Mathematics, Duke University

Department of Mathematics, University of Chicago

We study the homology of simplicial complexes built via deterministic rules from a random set of vertices. In particular, we show that, depending on the randomness that generates the vertices, the homology of these complexes can either become trivial as the number n of vertices grows, or can contain more and more complex structures. The different behaviours are consequences of different underlying distributions for the generation of vertices, and we consider three illustrative examples, when the vertices are sampled from Gaussian, exponential, and power-law distributions in \mathbb{R}^d .

We also discuss consequences of our results for manifold learning with noisy data, describing the topological phenomena that arise in this scenario as ‘crackle’, in analogy to audio crackle in temporal signal analysis.

1. Introduction. This paper treats the homology of simplicial complexes built via deterministic rules from a random set of vertices. In particular, it shows that, depending on the randomness that generates the vertices, the homology of these complexes can either become trivial as the sample size grows, or can contain more and more complex structures.

The motivation for these results comes from applications of topological tools for pattern analysis, object identification, and especially for the analysis of data sets. Typically, one starts with a collection of points and forms some simplicial complexes associated to these, and then takes their homology. For example, the 0-dimensional homology of such complexes can be interpreted as a version of clustering. The basic philosophy behind this attempt is that topology has an essentially qualitative nature and should therefore be robust with respect to small perturbations. Some recent references are [2, 3, 7, 13, 16] with two reviews, from different aspects, in [1] and [10]. Many of these papers find their *raison d’être* in essentially statistical problems, in which data generates the structures.

AMS 2000 subject classifications: Primary 60D05, 60F15, 60G55; Secondary 55U10.

Keywords and phrases: Čech complex, random complexes, persistent homology, random Betti numbers.

An important example occurs in the following manifold learning problem. Let \mathcal{M} be an unknown manifold embedded in a Euclidean space, and suppose that we are given a set of independent and identically distributed (i.i.d.) random samples $\mathcal{X}_n = \{X_1, \dots, X_n\}$ from the manifold. In order to recover the homology of \mathcal{M} , we consider the homology of

$$(1.1) \quad U = \bigcup_{k=1}^n B_\epsilon(X_k),$$

where $B_\epsilon(X)$ is the Euclidean ball, in the ambient space, of radius ϵ about the point X . The belief, or hope, is that, for large enough n , the homology of U will be equivalent to that of \mathcal{M} . A confounding issue arises when the sample points do not necessarily lie on the manifold, but rather are perturbed from it by a random amount. When this happens, it will follow from our results that the precise distribution behind the randomness plays a qualitatively important role. It is known that if the perturbations come from a bounded or strongly concentrated distribution, then they do not lead to much spurious homology, and the above line of attack, appropriately applied, works. For example, it was shown in [15] that for Gaussian noise it is possible to clean the data and recover the underlying topology of \mathcal{M} in a way that is essentially independent on the ambient dimension. Both [14, 15] contain results of the form that, given a nice enough \mathcal{M} , and any $\delta > 0$, there are explicit conditions on n and ϵ such that the homology of U is equal to the homology of \mathcal{M} with a probability of at least $(1 - \delta)$. However, for other distributions no such results exist, nor, in view of the results of this paper, are they to be expected.

Figure 1 provides an illustrative example of what happens when sampling points from an annulus and perturbing them with additional noise before reconstructing the annulus as in (1.1). In particular, it shows that if the additional noise is in some sense large then sample points can appear basically anywhere, introducing extraneous homology elements.

In order to be able, eventually, to extend the work in [15] beyond Gaussian noise, and make more concrete statements about the probabilistic features of the homology this extension generates, it is necessary to first focus on the behaviour of samples generated by pure noise, with no underlying manifold. In this case, thinking of the above setup, the manifold \mathcal{M} is simply the point at the origin, and the homology that we shall be trying to recapture is trivial. Nevertheless, we shall see that differing noise models can make this task extremely delicate, regardless of sample size.

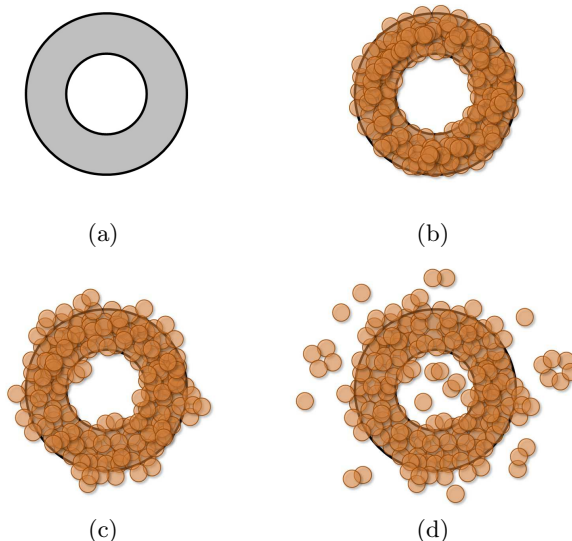


FIG 1. (a) The original space \mathcal{M} (an annulus) that we wish to recover from random samples. (b) With the appropriate choice of radius, we can easily recover the homology of the original space from random samples from \mathcal{M} . (c) In the presence of bounded noise, homology recovery is undamaged. (d) In the presence of unbounded noise, many extraneous homology elements appear, and significantly interfere with homology recovery.

1.1. *Some sample results.* To start being more concrete, let

$$\mathcal{X}_n = \{X_1, \dots, X_n\}$$

be a set of n i.i.d. random samples in \mathbb{R}^d , from a common density function f . Recall that the abstract simplicial complex $\check{C}(\mathcal{X}, \varepsilon)$ constructed according to the following rules is called the Čech complex associated to \mathcal{X} and ε :

1. The 0-simplices of $\check{C}(\mathcal{X}, \varepsilon)$ are the points in \mathcal{X} ,
2. An n -simplex $\sigma = [x_{i_0}, \dots, x_{i_n}]$ is in $\check{C}(\mathcal{X}, \varepsilon)$ if $\bigcap_{k=0}^n B_{x_{i_k}}(\varepsilon) \neq \emptyset$,

An important result, known as the ‘nerve theorem’, links Čech complexes and the neighborhood set U of (1.1), establishing that they are homotopy equivalent (cf. [5]). In particular, they have the same Betti numbers, measures of homology that we shall concentrate on in what follows.

If the sample distribution has a compact support S , then it is easy to show that, for large enough n ,

$$\check{C}(\mathcal{X}, \varepsilon) \simeq \bigcup_{k=1}^n B_\varepsilon(X_k) \approx \text{Tube}(S, \varepsilon) \triangleq \{x \in \mathbb{R}^d : \min_{y \in S} \|x - y\| \leq \varepsilon\},$$

where \simeq denotes homotopy equivalence and $\|\cdot\|$ is the standard L^2 norm in \mathbb{R}^d . Thus, there is not much to study in this case. However, when the support of the distribution is unbounded, interesting phenomena occur.

To study these phenomena, we shall consider three representative examples of probability densities. These are the *power-law*, *exponential*, and the *standard Gaussian* distributions, whose density functions are given, respectively, by

$$(1.2) \quad f_p(x) \triangleq \frac{c_p}{1 + \|x\|^\alpha},$$

$$(1.3) \quad f_e(x) \triangleq c_e e^{-\|x\|},$$

$$(1.4) \quad f_g(x) \triangleq c_g e^{-\|x\|^2/2},$$

where $\alpha > d$ and c_p, c_e, c_g are appropriate normalization constants that will not be of concern to us.

For large samples from any of these distributions we shall show that there exists a ‘core’ - a region in which the density of points is very high and so placing unit balls around them completely covers the region. Consequently, the Čech complex inside the core is contractible. The size of the core obviously grows to infinity as the sample size n goes to infinity, but its exact size will depend on the underlying distribution. For the three examples above, if we denote the radius of the core by R_n^c , we shall prove in Section 2.1 that

$$R_n^c \sim \begin{cases} (n/\log n)^{1/\alpha} & f(x) \propto \frac{1}{1+\|x\|^\alpha}, \\ \log n & f(x) \propto e^{-\|x\|}, \\ \sqrt{2 \log n} & f(x) \propto e^{-\|x\|^2/2}. \end{cases}$$

Note that in all three cases we have tacitly assumed that the cores are balls, a natural consequence of the spherical symmetry of the probability densities.

Beyond the core, the topology is more varied. For fixed n , there may be additional isolated components, but no longer enough placed densely enough to connect with one another and to form a contractible set. Indeed, we shall show that the individual components will typically have enough homology to be, individually, non-contractible. Thus, in this region, the topology of the Čech complex is highly nontrivial, and many homology elements of different orders appear. We call this phenomenon ‘crackling’, akin to the well known phenomenon caused by noise interference in audio signals and commonly referred to as crackling.

As for core size, the exact crackling behaviour depends on the choice of distribution. It turns out that Gaussian samples do not lead to crackling, but the other two cases do. To describe this, with some imprecision of notation

we shall write $[a, b)$ not only for an interval on the real line, but also for the annulus

$$[a, b) \triangleq \{x \in \mathbb{R}^d : a \leq \|x\| < b\}.$$

In Sections 2.2 and 2.3 we shall show that the exterior of the core can be divided into disjoint spherical annuli at radii

$$R_n^c \ll R_{d-1,n} \ll R_{d-2,n} \ll \cdots \ll R_{0,n}$$

(defined differently for each of the two crackling distributions) with different types of crackling (i.e. of homology) dominating in different regions.

In $[R_{0,n}, \infty)$ there are mostly disconnected points, and no structures with nontrivial homology. In $[R_{1,n}, R_{0,n})$ connectivity is a bit higher, and a finite number of 1-cycles appear. In $[R_{2,n}, R_{1,n})$ we have a finite number of 2-cycles, while the number of 1-cycles grows to infinity as $n \rightarrow \infty$. In general, in $[R_{k,n}, R_{k-1,n})$, as $n \rightarrow \infty$ we have a finite number of k -cycles, infinitely many l -cycles for $l < k$, and no cycles of dimension $l > k$. In other words, the crackle starts with a pure dust at $R_{n,0}$ and as we get closer to the core, higher dimensional homology gradually appears. See Figure 2 in the following section for more details.

As we already mentioned, the Gaussian distribution is fundamentally different than the other two, and does not lead to crackling. In Section 2.4 we show that, for the Gaussian distribution, there are hardly any points located outside the core. Thus, as $n \rightarrow \infty$, the union of balls around the sample points becomes a giant contractible ball of radius of order $\sqrt{2 \log n}$.

It is now possible to understand a little better how the results of this paper relate to the noisy manifold learning problem discussed above. For example, if the distribution of the noise is Gaussian, our results imply that if the manifold is well behaved, and the sample size is moderate, noise outliers should not significantly interfere with homology recovery, since Gaussian noise does not introduce artificial homology elements with large samples. However, there is a delicate counterbalance here between ‘moderate’ and ‘large’. Once the sample size is large, the core is also large, and the reconstructed manifold will have the topology of $\mathcal{M} \oplus B_{O(\sqrt{2 \log n})}(0)$, where \oplus is Minkowski addition. As n grows, the core will eventually envelope any compact manifold, and thus the homology of \mathcal{M} will be hidden by that of the core.

On the other hand, if the distribution of the noise is power-law or exponential, then noise outliers will typically generate extraneous homology elements that, for almost any sample size, will complicate the estimation of the original manifold. Furthermore, increasing the sample size in no way

solves this problem. Note that this issue is in addition to the fact that increasing the sample size will, as in the Gaussian case, create the problem of a large core concealing the topology of \mathcal{M} .

Thus, from a practical point of view, the message of this paper is that outliers cause problems in manifold estimation when noise is present, a fact well known to all practitioners who have worked in the area. What is qualitatively new here is a quantification of how this happens, and how it relates to the distribution of the noise. We do not attempt to solve this problem here, but unfortunately it follows from the results of this paper that algorithms for handling outliers will probably involve knowing at least the tail behaviour of the error distribution, despite the fact that in practical situations one does not generally want to take as known prior knowledge.

1.2. *On persistence intervals.* While the above discussion has concentrated on the persistence of noise induced crackle as sample sizes grow, and the regions in \mathbb{R}^d in which different types of homology appear, the proofs below also yield information about the more classical persistence diagrams of topological data analysis (cf. [6, 8–10]).

For example, in the two cases for which crackle persists – the power-law and exponential cases – estimates of the type appearing in Section 3 indicate that, with high probability, there exist extremely long bars in the bar code representation of persistent homology. Up to lower order corrections, preliminary calculations show that bar lengths for the k -th homology can be as large as $O(n^{\alpha_k})$ for the power-law case, and $\beta_k(\log \log n)$ for the exponential case, for appropriate α_k and β_k . More detailed studies of these phenomena will appear in a later publication.

1.3. *Poisson Processes.* Although we have described everything so far in terms of a random sample \mathcal{X} of n points taken from a density f , there is another way to approach the results of this paper, and that is to replace the points of \mathcal{X} with the points of a d -dimensional Poisson process \mathcal{P}_n whose intensity function is given by $\lambda_n = nf$. In this case the number of points is no longer fixed, but has mean n .

All the results of this paper stated for \mathcal{X} hold, without any change, if we replace \mathcal{X} by \mathcal{P} .

1.4. *Disclaimers.* Before starting the paper in earnest, and so as not to be accused of myopia, we note that the subject of manifold learning is obviously much broader than that described above, and algorithms for ‘estimating’ an underlying manifold from a finite sample abound in the statistics

and computer science literatures. Very few of them, however, take an algebraic point of view that we or the literature quoted above take. Furthermore, we note that other important results about the homology of Rips and Čech complexes for various distributions can be found in the papers [11, 12] and [4]. However, the methods and emphases of these papers are rather different.

2. Results. In this section we shall present all our main results, along with some discussion, more technical than that of the Introduction. Recall from Section 1.3 that although we present all results for the point set \mathcal{X} , they also hold if we replace the points of \mathcal{X} by the points of an appropriate Poisson process. All proofs are deferred to Section 3.

2.1. The Core of Distributions with Unbounded Support. We start by examining the core of the power-law, exponential and Gaussian distributions. These distributions are spherically symmetric and the samples are concentrated near the origin. By ‘core’ we refer to a centered ball $B_{R_n} \triangleq B_{R_n}(0) \subset \mathbb{R}^d$ containing a very large number of points from the sample \mathcal{X}_n , such that

$$B_{R_n} \subset \bigcup_{X \in \mathcal{X}_n \cap B_{R_n}} B_1(X).$$

i.e. the unit balls around the sample points completely cover B_{R_n} . In this case the homology of $\bigcup_{X \in \mathcal{X}_n \cap B_{R_n}} B_1(X)$, or equivalently, of $\check{C}(\mathcal{X}_n \cap B_{R_n}, 1)$, is trivial. Obviously, as $n \rightarrow \infty$, the radius R_n grows as well.

Let $\{R_n\}_{n=1}^\infty$ be an increasing sequence of positive numbers. Define by C_n the event that B_{R_n} is covered, i.e.

$$C_n \triangleq \left\{ B_{R_n} \subset \bigcup_{X \in \mathcal{X}_n \cap B_{R_n}} B_1(X) \right\}.$$

We wish to find the largest possible value of R_n such that $\mathbb{P}(C_n) \rightarrow 1$. The following theorem presents lower bounds for this value.

THEOREM 2.1. *Let $\epsilon > 0$, and define*

$$R_n^c \triangleq \begin{cases} \left(\frac{\delta_p n}{(\log n - e^{-\epsilon} \log \log n)} - 1 \right)^{1/\alpha} & f = f_p, \\ \log n - \log \log \log n - \delta_e - \epsilon & f = f_e, \\ \sqrt{2(\log n - \log \log \log n - \delta_g - \epsilon)} & f = f_g, \end{cases}$$

where the three distributions are given by (1.2)–(1.4), and

$$\begin{aligned}\delta_p &= c_p \alpha 2^{-d} d^{-(1+d/2)}, \\ \delta_e &= (1 + d/2) \log d + d \log 2 - \log c_e, \\ \delta_g &= (1 + d/2) \log d + (d - 1) \log 2 - \log c_g.\end{aligned}$$

If $R_n \leq R_n^c$, then

$$\mathbb{P}(C_n) \rightarrow 1.$$

Theorem 2.1 implies that the core size has a completely different order of magnitude for each of the three distributions. The heavy-tailed, power-law distribution has the largest core, while the core of the Gaussian distribution is the smallest. In the following sections we shall study the behaviour of the Čech complex outside the core.

2.2. How Power-Law Noise Crackles. In this section we explore the crackling phenomenon in the power-law distribution $f = f_p$. Let $B_{R_n} \subset \mathbb{R}^d$ be the centered ball with radius R_n , and let

$$\check{C}_n \triangleq \check{C}(\mathcal{X}_n \cap (B_{R_n})^c, 1),$$

be the Čech complex constructed from sample points outside B_{R_n} . We wish to study

$$\beta_{k,n} \triangleq \beta_k(\check{C}_n),$$

the k -th Betti number of \check{C}_n .

Note that the minimum number of points required to form a k -dimensional cycle ($k \geq 1$) is $k + 2$. For $k \geq 1$ and $\mathcal{Y} \subset \mathbb{R}^d$, denote

$$T_k(\mathcal{Y}) \triangleq \mathbb{1} \{ |\mathcal{Y}| = k + 2, \beta_k(\check{C}(\mathcal{Y}, 1)) = 1 \},$$

i.e. T_k takes the value 1 if $\check{C}(\mathcal{Y}, 1)$ is a minimal k -dimensional cycle, and 0 otherwise. This indicator function will be used to define the limits of the Betti numbers.

THEOREM 2.2. *If $\lim_{n \rightarrow \infty} n R_n^{-\alpha} = 0$, then*

$$\begin{aligned}\lim_{n \rightarrow \infty} \left(n R_n^{d-\alpha} \right)^{-1} \mathbb{E} \{ \beta_{0,n} \} &= \mu_{p,0}, \\ \lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-\alpha(k+2)} \right)^{-1} \mathbb{E} \{ \beta_{k,n} \} &= \mu_{p,k}, \quad 1 \leq k \leq d-1\end{aligned}$$

where

$$(2.1) \quad \mu_{p,0} \triangleq \frac{s_{d-1}c_p}{\alpha - d},$$

$$(2.2) \quad \mu_{p,k} \triangleq \frac{s_{d-1}c_p^{k+2}}{(\alpha(k+2) - d)(k+2)!} \int_{(\mathbb{R}^d)^{k+1}} T_k(0, \mathbf{y}) d\mathbf{y}, \quad 1 \leq k \leq d-1,$$

and where s_{d-1} is the surface area of the $(d-1)$ -dimensional unit sphere in \mathbb{R}^d .

Next, we define the following values, which will serve as critical radii for the crackle,

$$\begin{aligned} R_{0,n}^\epsilon &\triangleq n^{\left(\frac{1}{\alpha-d} + \epsilon\right)}, \\ R_{0,n} &\triangleq R_{0,n}^0, \\ R_{k,n}^\epsilon &\triangleq n^{\left(\frac{1}{\alpha-d/(k+2)} + \epsilon\right)}, \quad (k \geq 1) \\ R_{k,n} &\triangleq R_{k,n}^0. \end{aligned}$$

The following is a straightforward corollary of Theorem 2.2, and summarizes the behaviour of $\mathbb{E}\{\beta_{k,n}\}$ in the power-law case.

COROLLARY 2.3. For $k \geq 0$ and $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{E}\{\beta_{k,n}\} = \begin{cases} 0 & R_n = R_{k,n}^\epsilon, \\ \mu_{p,k} & R_n = R_{k,n}, \\ \infty & R_n = R_{k,n}^{-\epsilon}, \end{cases}$$

Theorem 2.2 and Corollary 2.3 reveal that the crackling behaviour is organized into separate ‘layers’, see Figure 2. Dividing \mathbb{R}^d into a sequence of annuli at radii

$$R_{0,n}^\epsilon \gg R_{0,n} \gg R_{1,n}^\epsilon \gg R_{1,n} \gg \cdots \gg R_{d-1,n}^\epsilon \gg R_{d-1,n} \gg R_n^c,$$

we observe a different behaviour of the Betti numbers in each annulus. We shall briefly review the behaviour in each annulus, in a decreasing order of radii values. The following description is mainly qualitative, and refers to expected values only.

- $[R_{0,n}^\epsilon, \infty)$ - there are hardly any points ($\beta_k \sim 0$, $0 \leq k \leq d-1$).
- $[R_{0,n}, R_{0,n}^\epsilon)$ - points start to appear, and $\beta_0 \sim \mu_{p,0}$. The points are very few and scattered, so no cycles are generated ($\beta_k \sim 0$, $1 \leq k \leq d-1$).

- $[R_{1,n}^\epsilon, R_{0,n})$ - the number of components grows to infinity, but no cycles are formed yet ($\beta_0 \sim \infty$, and $\beta_k = 0$, $1 \leq k \leq d-1$).
- $[R_{1,n}, R_{1,n}^\epsilon)$ - a finite number of 1-dimensional cycles show up, among the infinite number of components ($\beta_0 \sim \infty$, $\beta_1 \sim \mu_{p,1}$, and $\beta_k = 0$, $1 \leq k \leq d-1$).
- $[R_{2,n}^\epsilon, R_{1,n})$ - we have $\beta_0 \sim \infty$, $\beta_1 \sim \infty$, and $\beta_k \sim 0$ for $k \geq 1$.

This process goes on, until the $(d-1)$ -dimensional cycles appear -

- $[R_{d-1}, R_{d-1}^\epsilon)$ - we have $\beta_{d-1} \sim \mu_{p,d-1}$ and $\beta_k \sim \infty$ for $0 \leq k \leq d-2$.
- $[R_n^c, R_{d-1})$ - just before we reach the core, the complex exhibits the most intricate structure, with $\beta_k \sim \infty$ for $0 \leq k \leq d-1$.

Note that there is a very fast phase transition as we move from the contractible core to the first crackle layer. At this point we do not know exactly where and how this phase transition takes place. A reasonable conjecture would be that the transition occurs at $R_n = n^{1/\alpha}$ (since at this radius the term $nR_n^{-\alpha}$ that appears in Theorem 2.2 changes its limit, affecting the limiting Betti numbers). However, this remains for future work.

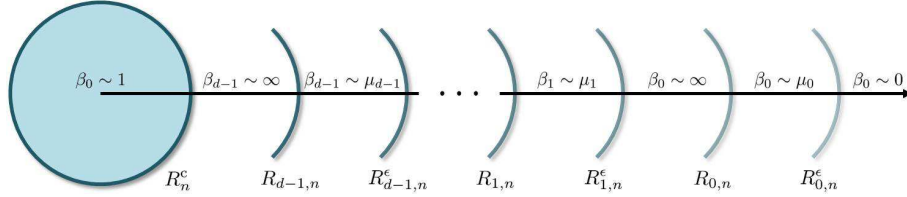


FIG 2. *The layered behaviour of crackle. Inside the core ($B_{R_n^c}$) the complex consists of a single component and no cycles. The exterior of the core is divided into separate annuli. Going from right to left, we see how the Betti numbers grow. In each annulus we present the Betti number that was most recently changed.*

2.3. How Exponential Noise Crackles. In this section we focus on the exponential density function $f = f_e$. The results in this section are very similar to the those for the power law distribution, and we shall describe them briefly. Differences lie in the specific values of the $R_{k,n}$ and in the terms in the limit formulae.

THEOREM 2.4. *If $\lim_{n \rightarrow \infty} ne^{-R_n} = 0$, then*

$$\lim_{n \rightarrow \infty} \left(nR_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{ \beta_{0,n} \} = \mu_{e,0},$$

$$\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E} \{ \beta_{k,n} \} = \mu_{e,k}, \quad k \geq 1$$

where

(2.3)

$$\mu_{e,0} \triangleq s_{d-1} c_e,$$

(2.4)

$$\mu_{e,k} \triangleq \frac{s_{d-1} c_e^{k+2}}{(k+2)!} \int_0^\infty \int_{(\mathbb{R}^d)^{k+1}} T_k(0, \mathbf{y}) e^{-(k+2)\rho + \sum_{i=1}^{k+1} y_i^1} \prod_{i=1}^{k+1} \mathbb{1}\{y_i^1 > -\rho\} d\mathbf{y} d\rho,$$

and where y_i^1 is the first coordinate of $y_i \in \mathbb{R}^d$.

Next, define

$$\begin{aligned} R_{0,n}^\epsilon &\triangleq \log n + (d-1+\epsilon) \log \log n, \\ R_{0,n} &\triangleq R_{0,n}^0, \\ R_{k,n}^\epsilon &\triangleq \log n + \left(\frac{d-1}{k+2} + \epsilon \right) \log \log n, \quad (k \geq 1) \\ R_{k,n} &\triangleq R_{k,n}^0. \end{aligned}$$

From Theorem 2.4 we can conclude the following.

COROLLARY 2.5. For $k \geq 0$ and $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{E}\{\beta_{k,n}\} = \begin{cases} 0 & R_n = R_{k,n}^\epsilon, \\ \mu_{e,k} & R_n = R_{k,n}, \\ \infty & R_n = R_{k,n}^{-\epsilon}, \end{cases}$$

As in the power-law case, Theorem 2.4 implies the same ‘layered’ behaviour, the only difference being in the values of $R_{k,n}$. From examining the values of R_n^c , and $R_{k,n}$ it is reasonable to guess that the phase transition in the exponential case occurs at $R_n = \log n$.

2.4. Gaussian Noise Does Not Crackle. Simplicial complexes built over vertices sampled from the standard Gaussian distribution exhibit a completely different behaviour to that we saw in the power-law and exponential cases. Define

$$R_{0,n}^\epsilon \triangleq \sqrt{2 \log n + (d-2+\epsilon) \log \log n},$$

then

THEOREM 2.6. *If $f = f_g$, $\epsilon > 0$, and $R_n = R_{0,n}^\epsilon$, then for $0 \leq k \leq d-1$*

$$\lim_{n \rightarrow \infty} \mathbb{E} \{\beta_{k,n}\} = 0.$$

Note that in the Gaussian case $\lim_{n \rightarrow \infty} (R_{0,n}^\epsilon - R_n^c) = 0$. This implies that as $n \rightarrow \infty$ we have the core which is contractible, and outside the core there is hardly anything. In other words, the ball placed around every new point we add to the sample immediately connects to the core, and thus, the Gaussian noise *does not crackle*.

3. Proofs. We now turn to proofs, starting with the proof of the main result of Section 2.1.

3.1. The Core.

PROOF OF THEOREM 2.1. The proof covers all three distributions, except for specific calculations near the end. Take a grid on \mathbb{R}^d of size $g = \frac{1}{2\sqrt{d}}$. Let \mathcal{Q}_n be the collection of cubes in this grid that are contained in B_{R_n} . Let \tilde{C}_n be the following event

$$\tilde{C}_n \triangleq \{\forall Q \in \mathcal{Q}_n : Q \cap \mathcal{X}_n \neq \emptyset\},$$

i.e. \tilde{C}_n is the event that every cube in \mathcal{Q}_n contains at least one point from \mathcal{X}_n . Recall the definition of C_n ,

$$C_n \triangleq \left\{ B_{R_n} \subset \bigcup_{X \in \mathcal{X}_n \cap B_{R_n}} B_1(X) \right\}.$$

Then it is easy to show that $\tilde{C}_n \subset C_n$. The complementary event \tilde{C}_n^c is the event that at least one cube is empty. Thus,

$$\mathbb{P}(\tilde{C}_n^c) \leq \sum_{Q \in \mathcal{Q}_n} \mathbb{P}(Q \cap \mathcal{X}_n = \emptyset) = \sum_{Q \in \mathcal{Q}_n} (1 - p(Q))^n \leq \sum_{Q \in \mathcal{Q}_n} e^{-np(Q)}$$

where

$$p(Q) = \int_Q f(z) dz \geq g^d f(R_n).$$

In addition, the number of cubes that are contained in B_{R_n} is less than $(2R_n/g)^d$. Therefore,

$$(3.1) \quad \mathbb{P}(\tilde{C}_n^c) \leq (2g^{-1})^d R_n^d e^{-ng^d f(R_n)}.$$

Now, choose any $\epsilon > 0$ and set

$$R_n = R_n^c \triangleq \begin{cases} \left(\frac{\delta_p n}{(\log n - e^{-\epsilon} \log \log n)} - 1 \right)^{1/\alpha} & f = f_p, \\ \log n - \log \log \log n - \delta_e - \epsilon & f = f_e, \\ \sqrt{2} (\log n - \log \log \log n - \delta_g - \epsilon) & f = f_g, \end{cases}$$

where

$$\begin{aligned} \delta_p &= c_p \alpha 2^{-d} d^{-(1+d/2)}, \\ \delta_e &= \log d - \log c_e - \log g^d, \\ \delta_g &= \log(d/2) - \log c_g - \log g^d. \end{aligned}$$

It is easy to verify that in all cases we have

$$R_n^d e^{-ng^d f(R_n)} \rightarrow 0.$$

Thus, from (3.1) we conclude that $\mathbb{P}(\tilde{C}_n) \rightarrow 1$. Since $\mathbb{P}(C_n) \geq \mathbb{P}(\tilde{C}_n)$ we now have that for $R_n = R_n^c$, in each of the distributions,

$$\mathbb{P}(C_n) \rightarrow 1,$$

which completes the proof. \square

3.2. Crackle - Notation and General Lemmas. For $R_n > 0$, set

$$\mathcal{X}_{n,R_n} \triangleq \mathcal{X}_n \cap (B_{R_n})^c,$$

i.e. \mathcal{X}_{n,R_n} consists of the points of \mathcal{X}_n located outside the ball B_{R_n} . Next, recall the definition of T_k ,

$$T_k(\mathcal{Y}) \triangleq \mathbb{1} \{ |\mathcal{Y}| = k + 2, \beta_k(\check{C}(\mathcal{Y}, 1)) = 1 \},$$

for $\mathcal{Y} \subset \mathbb{R}^d$, and write

$$\begin{aligned} S_{0,n} &\triangleq |\mathcal{X}_{n,R_n}|, \\ \hat{S}_{0,n} &\triangleq \# \{ X \in \mathcal{X}_{n,R_n} : X \text{ is a connected component of } \check{C}(\mathcal{X}_n, 1) \} \\ S_{k,n} &\triangleq \sum_{\mathcal{Y} \subset \mathcal{X}_{n,R_n}} T_k(\mathcal{Y}), \\ \hat{S}_{k,n} &\triangleq \sum_{\mathcal{Y} \subset \mathcal{X}_{n,R_n}} T_k(\mathcal{Y}) \mathbb{1} \{ \check{C}(\mathcal{Y}, 1) \text{ is a connected component of } \check{C}(\mathcal{X}_n, 1) \}, \\ L_{k,n} &\triangleq \sum_{\mathcal{Y} \subset \mathcal{X}_{n,R_n}} \mathbb{1} \{ |\mathcal{Y}| = k + 3, \check{C}(\mathcal{Y}, 1) \text{ is connected} \}, \end{aligned}$$

where $k \geq 1$. Observe that

$$(3.2) \quad \hat{S}_{0,n} \leq \beta_{0,n} \leq S_{0,n}$$

$$(3.3) \quad \hat{S}_{k,n} \leq \beta_{k,n} \leq \hat{S}_{k,n} + L_{k,n}, \quad k \geq 1$$

We will evaluate the limits of $\mathbb{E}\{S_{k,n}\}$, $\mathbb{E}\{\hat{S}_{k,n}\}$ and $\mathbb{E}\{L_{k,n}\}$ and deduce from these the limit of $\mathbb{E}\{\beta_{k,n}\}$.

In addition, set

$$\begin{aligned} \mathbf{e}_1 &\triangleq (1, 0, \dots, 0) \in \mathbb{R}^d, \\ f(r) &\triangleq f(r\mathbf{e}_1), \quad r \in \mathbb{R}, \\ U(\mathbf{x}) &\triangleq \bigcup_{i=1}^k B_2(x_i), \quad \mathbf{x} \in (\mathbb{R}^d)^k, \\ p(\mathbf{x}) &\triangleq \int_{U(\mathbf{x})} f(z) dz, \quad \mathbf{x} \in (\mathbb{R}^d)^k. \end{aligned}$$

The following two lemmas are purely technical, but will considerably simplify our computations later.

LEMMA 3.1. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a spherically symmetric probability density. Then,*

$$\begin{aligned} \mathbb{E}\{S_{0,n}\} &= s_{d-1}n \int_{R_n}^{\infty} r^{d-1} f(r) dr, \\ \mathbb{E}\{\hat{S}_{0,n}\} &= s_{d-1}n \int_{R_n}^{\infty} r^{d-1} f(r) (1 - np(r\mathbf{e}_1))^{n-1} dr, \end{aligned}$$

where s_{d-1} is the volume of the $d-1$ dimensional unit sphere.

PROOF. $S_{0,n}$ is simply a sum of Bernoulli variables, therefore

$$\mathbb{E}\{S_{0,n}\} = n\mathbb{P}(\|X\| > R_n) = n \int_{\mathbb{R}^d} f(x) \mathbb{1}_{\{\|x\| > R_n\}} dx.$$

Writing the integral in polar coordinates yields

$$\mathbb{E}\{S_{0,n}\} = n \int_{R_n}^{\infty} \int_{S^{d-1}} f(r\theta) r^{d-1} J(\theta) d\theta dr,$$

where $J(\theta) = \left| \frac{\partial x}{\partial \theta} \right|$. Since f is spherically symmetric, $f(r\theta) = f(r)$, and therefore

$$\mathbb{E}\{S_{0,n}\} = s_{d-1}n \int_{R_n}^{\infty} r^{d-1} f(r) dr.$$

The proof for $\hat{S}_{0,n}$ is similar, using the fact that the probability that a point $x \in \mathbb{R}^d$ is disconnected from the rest of the complex $\tilde{C}(\mathcal{X}_n, 1)$ is $(1-p(x))^{n-1}$. \square

LEMMA 3.2. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a spherically symmetric probability density. Then, for $k \geq 1$,*

$$\begin{aligned}\mathbb{E}\{S_{k,n}\} &= s_{d-1} \binom{n}{k+2} \int_{R_n}^{\infty} r^{d-1} f(r) G_k(r) dr, \\ \mathbb{E}\{\hat{S}_{k,n}\} &= s_{d-1} \binom{n}{k+2} \int_{R_n}^{\infty} r^{d-1} f(r) \hat{G}_k(r) dr,\end{aligned}$$

where s_{d-1} is the volume of the $d-1$ dimensional sphere, and where

$$\begin{aligned}G_k(r) &\triangleq \int_{(\mathbb{R}^d)^{k+1}} f(\|re_1 + \mathbf{y}\|) T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}\{\|re_1 + y_i\| > R_n\} d\mathbf{y}, \\ \hat{G}_k(r) &\triangleq \int_{(\mathbb{R}^d)^{k+1}} f(\|re_1 + \mathbf{y}\|) T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}\{\|re_1 + y_i\| > R_n\} \\ &\quad \times (1 - p(re_1, re_1 + \mathbf{y}))^{n-k-2} d\mathbf{y}.\end{aligned}$$

PROOF. The proof is in the same spirit of the proof of Lemma 3.1, but technically more complicated. Thinking of $S_{k,n}$ as a sum of Bernoulli variables, we have that

$$\mathbb{E}\{S_{k,n}\} = \binom{n}{k+2} \int_{(\mathbb{R}^d)^{k+2}} f(\mathbf{x}) T_k(\mathbf{x}) \prod_{i=1}^{k+2} \mathbb{1}\{\|x_i\| > R_n\} d\mathbf{x}.$$

Let I_k denote the integral above. Then, using the change of variables

$$x_1 \rightarrow x, \quad x_i \rightarrow x + y_{i-1}, \quad (i > 1),$$

yields

$$\begin{aligned}I_k &= \int_{\|x\| \geq R_n} \int_{(\mathbb{R}^d)^{k+1}} f(x) f(x + \mathbf{y}) T_k(x, x + \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}\{\|x + y_i\| > R_n\} d\mathbf{y} dx \\ &= \int_{\|x\| \geq R_n} \int_{(\mathbb{R}^d)^{k+1}} f(x) f(x + \mathbf{y}) T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}\{\|x + y_i\| > R_n\} d\mathbf{y} dx.\end{aligned}$$

Moving to polar coordinates yields

$$\begin{aligned}
I_k &= \int_{R_n}^{\infty} \int_{S^{d-1}} \int_{(\mathbb{R}^d)^{k+1}} f(r\theta) f(r\theta + \mathbf{y}) T_k(0, \mathbf{y}) \\
&\quad \times \prod_{i=1}^{k+1} \mathbb{1}_{\{\|r\theta + y_i\| > R_n\}} r^{d-1} J(\theta) d\mathbf{y} d\theta dr \\
&= \int_{R_n}^{\infty} r^{d-1} f(r) \int_{S^{d-1}} J(\theta) \int_{(\mathbb{R}^d)^{k+1}} f(\|r\theta + \mathbf{y}\|) T_k(0, \mathbf{y}) \\
&\quad \times \prod_{i=1}^{k+1} \mathbb{1}_{\{\|r\theta + y_i\| > R_n\}} d\mathbf{y} d\theta dr,
\end{aligned}$$

where $J(\theta) = \left| \frac{\partial x}{\partial \theta} \right|$, and $f(x) = f(\|x\|)$ by the spherical symmetry assumption. Set

$$G_k(r, \theta) \triangleq \int_{(\mathbb{R}^d)^{k+1}} f(\|r\theta + \mathbf{y}\|) T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}_{\{\|r\theta + y_i\| > R_n\}} d\mathbf{y}.$$

Since T_k is rotation invariant, it is easy to show that for every $\theta \in S^{d-1}$

$$G_k(r, \theta) = G_k(r, \mathbf{e}_1) \triangleq G_k(r).$$

Thus,

$$(3.4) \quad I_k = s_{d-1} \int_{R_n}^{\infty} r^{d-1} f(r) G_k(r) dr.$$

This completes the proof for $S_{k,n}$. The proof for $\widehat{S}_{k,n}$ is similar. □

In what follows, we shall use the following elementary limits:

1. For every $k > 0$,

$$(3.5) \quad \lim_{n \rightarrow \infty} n^{-k} \binom{n}{k} = \frac{1}{k!}$$

2. For every sequence $a_n \rightarrow 0$ and $k \geq 0$,

$$(3.6) \quad \lim_{n \rightarrow \infty} \frac{(1 - a_n)^{n-k}}{e^{-na_n}} = 1$$

3.3. *Crackle - The Power Law Distribution.* In this section we prove the results in Section 2.2. First, we need a few lemmas.

LEMMA 3.3. *If $f = f_p$, and $R_n \rightarrow \infty$, then*

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-\alpha} \right)^{-1} \mathbb{E} \{ S_{0,n} \} = \mu_{p,0},$$

where $\mu_{p,0}$ is defined in (2.1).

If, in addition, $n R_n^{-\alpha} \rightarrow 0$, then

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-\alpha} \right)^{-1} \mathbb{E} \{ \hat{S}_{0,n} \} = \mu_{p,0}.$$

PROOF. From Lemma 3.1 we have that

$$\mathbb{E} \{ S_{0,n} \} = s_{d-1} n \int_{R_n}^{\infty} r^{d-1} f(r) dr.$$

Making the change of variables $r \rightarrow R_n \rho$ yields

$$\begin{aligned} \mathbb{E} \{ S_{0,n} \} &= s_{d-1} n \int_1^{\infty} \frac{c_p (R_n \rho)^{d-1}}{1 + (R_n \rho)^\alpha} R_n d\rho \\ &= s_{d-1} c_p n R_n^{d-\alpha} \int_1^{\infty} \frac{\rho^{d-1}}{R_n^{-\alpha} + \rho^\alpha} d\rho. \end{aligned}$$

Applying the dominated convergence theorem to the previous integral gives

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-\alpha} \right)^{-1} \mathbb{E} \{ S_{0,n} \} = s_{d-1} c_p \int_1^{\infty} \rho^{d-1-\alpha} d\rho = \frac{s_{d-1} c_p}{\alpha - d} = \mu_{p,0}.$$

This proves the first part of the lemma.

Next, from Lemma 3.1 we have that

$$\mathbb{E} \{ \hat{S}_{0,n} \} = s_{d-1} n \int_{R_n}^{\infty} r^{d-1} f(r) (1 - p(re_1))^{n-1} dr.$$

The power term is bounded by 1 and therefore will not affect the conditions needed for dominated convergence. Thus, using (3.6), we only need to evaluate its limit.

$$p(re_1) = \int_{B_2(re_1)} f(z) dz = \int_{B_2(0)} \frac{c_p}{1 + \|re_1 + z\|} dz,$$

and after the change of variables $r \rightarrow R_n \rho$ we have,

$$p(R_n \rho e_1) = c_p R_n^{-\alpha} \int_{B_2(0)} \frac{1}{R_n^{-\alpha} + \|\rho e_1 + R_n^{-1} z\|^\alpha} dz.$$

If $nR_n^{-\alpha} \rightarrow 0$, then, by dominated convergence, we have

$$\lim_{n \rightarrow \infty} np(R_n \rho e_1) = 0.$$

Thus,

$$\lim_{n \rightarrow \infty} (1 - p(R_n \rho e_1))^{n-1} = \lim_{n \rightarrow \infty} e^{-np(R_n \rho e_1)} = 1,$$

and therefore we have

$$\lim_{n \rightarrow \infty} \left(nR_n^{d-\alpha} \right)^{-1} \mathbb{E}\{\hat{S}_{0,n}\} = \lim_{n \rightarrow \infty} \left(nR_n^{d-\alpha} \right)^{-1} \mathbb{E}\{S_{0,n}\} = \mu_{p,0}.$$

This completes the proof of the second part of the lemma. \square

LEMMA 3.4. *If $f = f_p$, and $R_n \rightarrow \infty$ then*

$$\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-\alpha(k+2)} \right)^{-1} \mathbb{E}\{S_{k,n}\} = \mu_{p,k},$$

where $\mu_{p,k}$ is defined in (2.2). If, in addition, $nR_n^{-\alpha} \rightarrow 0$, then

$$\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-\alpha(k+2)} \right)^{-1} \mathbb{E}\{\hat{S}_{k,n}\} = \mu_{p,k}.$$

PROOF. The proof is in the spirit of the proof of Lemma 3.3, but technically more complicated. From Lemma 3.2 we have that

$$\mathbb{E}\{S_{k,n}\} = \binom{n}{k+2} I_k,$$

where

$$I_k = s_{d-1} \int_{R_n}^{\infty} r^{d-1} f(r) G_k(r) dr.$$

Making the change of variables $r \rightarrow R_n \rho$ yields

$$\begin{aligned} I_k &= s_{d-1} R_n \int_1^{\infty} (R_n \rho)^{d-1} f(R_n \rho) G_k(R_n \rho) d\rho \\ &= s_{d-1} c_p^{k+2} (R_n)^{d-\alpha(k+2)} \int_1^{\infty} \int_{(\mathbb{R}^d)^{k+1}} \frac{\rho^{d-1}}{R_n^{-\alpha} + \rho^{\alpha}} \prod_{i=1}^{k+1} \frac{1}{R_n^{-\alpha} + \|\rho e_1 + R_n^{-1} y_i\|^{\alpha}} \\ &\quad \times T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1}\{\|\rho e_1 + R_n^{-1} y_i\| > 1\} d\mathbf{y}. \end{aligned}$$

Thus, using (3.5),

$$\begin{aligned} (n^{k+2} R_n^{d-\alpha(k+2)})^{-1} \mathbb{E} \{S_{k,n}\} &= \frac{s_{d-1} c_p^{k+2}}{(k+2)!} \int_1^\infty \int_{(\mathbb{R}^d)^{k+1}} \frac{\rho^{d-1}}{R_n^{-\alpha} + \rho^\alpha} \\ &\times T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \frac{1}{R_n^{-\alpha} + \|\rho \mathbf{e}_1 + R_n^{-1} y_i\|^\alpha} \mathbb{1} \{ \|\rho \mathbf{e}_1 + R_n^{-1} y_i\| > 1 \} d\mathbf{y}. \end{aligned}$$

It is easy to show that the integrand is bounded by an integrable term, so the dominated convergence theorem applies, yielding

$$\begin{aligned} \lim_{n \rightarrow \infty} (n^{k+2} R_n^{d-\alpha(k+2)})^{-1} \mathbb{E} \{S_{k,n}\} &= \frac{s_{d-1} c_p^{k+2}}{(k+2)!} \int_1^\infty \rho^{d-1-\alpha(k+2)} d\rho \int_{(\mathbb{R}^d)^{k+1}} T_k(0, \mathbf{y}) d\mathbf{y} \\ &= \frac{s_{d-1} c_p^{k+2}}{(\alpha(k+2) - d)(k+2)!} \int_{(\mathbb{R}^d)^{k+1}} T_k(0, \mathbf{y}) d\mathbf{y} \\ &= \mu_{p,k}. \end{aligned}$$

This proves the first part of the lemma.

Next, the terms $G_k(r)$ and $\hat{G}_k(r)$ in Lemma 3.2 differ only by the term $(1 - p(re_1, re_1 + \mathbf{y}))^{n-k-2}$, so dominated convergence still applies. Now,

$$p(re_1, re_1 + \mathbf{y}) = \int_{U(re_1, re_1 + \mathbf{y})} f(z) dz = \int_{U(0, \mathbf{y})} f(re_1 + z) dz,$$

and substituting $r \rightarrow R_n \rho$ yields,

$$p(R_n \rho e_1, R_n \rho e_1 + \mathbf{y}) = c_p R_n^{-\alpha} \int_{U(0, \mathbf{y})} \frac{1}{R_n^{-\alpha} + \|\rho e_1 + R_n^{-1} z\|^\alpha} dz.$$

If $n R_n^{-\alpha} \rightarrow 0$, then using the dominated convergence we have

$$\lim_{n \rightarrow \infty} np(R_n \rho e_1, R_n \rho e_1 + \mathbf{y}) = 0.$$

Thus,

$$\lim_{n \rightarrow \infty} e^{-np(R_n \rho e_1, R_n \rho e_1 + \mathbf{y})} = 1,$$

and therefore, using (3.6),

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-\alpha(k+2)} \right)^{-1} \mathbb{E} \{ \hat{S}_{k,n} \} &= \lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-\alpha(k+2)} \right)^{-1} \mathbb{E} \{ S_{k,n} \} \\ &= \mu_{p,k}. \end{aligned}$$

This completes the proof of the second part of the lemma. \square

LEMMA 3.5. *If $f = f_p$, and $R_n \rightarrow \infty$ then*

$$\lim_{n \rightarrow \infty} \left(n^{k+3} R_n^{d-\alpha(k+3)} \right)^{-1} \mathbb{E} \{L_{k,n}\} = \hat{\mu}_{p,k},$$

for some $\hat{\mu}_{p,k} > 0$.

PROOF. The proof is very similar to the proof of Lemma 3.4. We need only replace T_k with an indicator function that tests whether a sub-complex generated by $k+3$ points is connected. The exact value of $\hat{\mu}_{p,k}$ will not be needed anywhere. \square

We can now prove Theorem 2.2.

PROOF OF THEOREM 2.2. To prove the limit for $\beta_{0,n}$ simply combine Lemma 3.3 with the inequality (3.2). To prove the limit for $\beta_{k,n}$, $k \geq 1$, combine Lemmas 3.4 and 3.5 with the inequality (3.3). \square

3.4. *Crackle - The Exponential Distribution.* In this section we wish to prove Theorem 2.4. We start with the following lemmas.

LEMMA 3.6. *If $f = f_e$, and $R_n \rightarrow \infty$ then,*

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{S_{0,n}\} = \mu_{e,0},$$

where $\mu_{e,0}$ is defined in (2.3).

If, in addition, $n e^{-R_n} \rightarrow 0$ then,

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{\hat{S}_{0,n}\} = \mu_{e,0}.$$

PROOF. From Lemma 3.1 we have that

$$\mathbb{E} \{S_{0,n}\} = s_{d-1} n \int_{R_n}^{\infty} r^{d-1} f(r) dr.$$

Using the change of variables $r \rightarrow \rho + R_n$ yields

$$\begin{aligned} \mathbb{E} \{S_{0,n}\} &= s_{d-1} n \int_0^{\infty} (\rho + R_n)^{d-1} c_e e^{-(\rho+R_n)} d\rho \\ &= s_{d-1} c_e n R_n^{d-1} e^{-R_n} \int_0^{\infty} \left(\frac{\rho}{R_n} + 1 \right)^{d-1} e^{-\rho} d\rho. \end{aligned}$$

Applying dominated convergence to the last integral yields,

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{ S_{0,n} \} = s_{d-1} c_e \int_0^\infty e^{-\rho} d\rho = s_{d-1} c_e = \mu_{e,0}.$$

This proves the first part of the lemma.

Next, from Lemma 3.1 we have that

$$\mathbb{E} \{ \hat{S}_{0,n} \} = s_{d-1} n \int_{R_n}^\infty r^{d-1} f(r) (1 - p(re_1))^{n-1} dr.$$

The power term will not affect the dominated convergence conditions. Thus, we only need to evaluate its limit.

$$p(re_1) = \int_{B_2(re_1)} f(z) dz = \int_{B_2(0)} c_e e^{-\|re_1+z\|} dz,$$

and after the change of variables $r \rightarrow \rho + R_n$ we have,

$$p((\rho + R_n)e_1) = \int_{B_2(0)} c_e e^{-\|(\rho+R_n)e_1+z\|} dz \leq e^{-(R_n+\rho)} \int_{B_2(0)} c_e e^{\|z\|} dz.$$

If $ne^{-R_n} \rightarrow 0$, then

$$\lim_{n \rightarrow \infty} np((\rho + R_n)e_1) = 0.$$

Thus,

$$\lim_{n \rightarrow \infty} e^{-np((\rho+R_n)e_1)} = 1,$$

and therefore, using (3.6), we have

$$\lim_{n \rightarrow \infty} \left(n R_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{ \hat{S}_{0,n} \} = \lim_{n \rightarrow \infty} \left(n R_n^{d-1} e^{-R_n} \right)^{-1} \mathbb{E} \{ S_{0,n} \} = \mu_{e,0}.$$

This completes the proof of the second part of the lemma. \square

LEMMA 3.7. *If $f = f_e$, and $R_n \rightarrow \infty$ then,*

$$\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E} \{ S_{k,n} \} = \mu_{e,k},$$

where $\mu_{e,k}$ is defined in (2.4).

If, in addition, $ne^{-R_n} \rightarrow 0$ then,

$$\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E} \{ \hat{S}_{k,n} \} = \mu_{e,k}.$$

PROOF. From Lemma 3.2 we have that

$$\mathbb{E} \{S_{k,n}\} = \frac{n^{k+2}}{(k+2)!} I_k,$$

where

$$I_k = s_{d-1} \int_{R_n}^{\infty} r^{d-1} f(r) G_k(r) dr.$$

Making the change of variables $r \rightarrow \rho + R_n$ yields

$$\begin{aligned} I_k &= s_{d-1} \int_0^{\infty} (\rho + R_n)^{d-1} f(\rho + R_n) G_k(\rho + R_n) d\rho \\ &= s_{d-1} c_e^{k+2} \int_0^{\infty} \int_{(\mathbb{R}^d)^{k+1}} (\rho + R_n)^{d-1} e^{-(\rho+R_n)} \prod_{i=1}^{k+1} e^{-\|(\rho+R_n)\mathbf{e}_1 + y_i\|} \\ &\quad \times T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} \mathbb{1} \{ \|(\rho + R_n)\mathbf{e}_1 + y_i\| > R_n \} d\mathbf{y} d\rho \\ &= s_{d-1} c_e^{k+2} e^{-(k+2)R_n} R_n^{d-1} \int_0^{\infty} \int_{(\mathbb{R}^d)^{k+1}} \left(\frac{\rho}{R_n} + 1 \right)^{d-1} e^{-\rho} \\ &\quad \times T_k(0, \mathbf{y}) \prod_{i=1}^{k+1} e^{-\|(\rho+R_n)\mathbf{e}_1 + y_i\|} e^{R_n} \mathbb{1} \{ \|(\rho + R_n)\mathbf{e}_1 + y_i\| > R_n \} d\mathbf{y} d\rho. \end{aligned}$$

The last integral can be easily shown to satisfy the conditions of the dominated convergence theorem. In addition, it is easy to show that

$$\lim_{n \rightarrow \infty} e^{-\|(\rho+R_n)\mathbf{e}_1 + y_i\|} e^{R_n} = e^{-(\rho + \langle \mathbf{e}_1, y_i \rangle)} = e^{-(\rho + y_i^1)},$$

where y_i^1 is the first coordinate of $y_i \in \mathbb{R}^d$, and also that

$$\lim_{n \rightarrow \infty} \mathbb{1} \{ \|(\rho + R_n)\mathbf{e}_1 + y_i\| > R_n \} = \mathbb{1} \{ y_i^1 \geq -\rho \}.$$

Altogether, we have that

$$\begin{aligned} &\lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E} \{S_{k,n}\} \\ &= \frac{s_{d-1} c_e^{k+2}}{(k+2)!} \int_0^{\infty} \int_{(\mathbb{R}^d)^{k+1}} T_k(0, \mathbf{y}) e^{-((k+2)\rho + \sum_{i=1}^{k+1} y_i^1)} \prod_{i=1}^{k+1} \mathbb{1} \{ y_i^1 \geq -\rho \} d\mathbf{y} d\rho, \end{aligned}$$

proving the first part of the lemma.

Next, as in the proof of Lemma 3.4, we need to evaluate the term $p(re_1, re_1 + \mathbf{y})$.

$$p(re_1, re_1 + \mathbf{y}) = \int_{U(0, \mathbf{y})} c_e e^{-\|re_1 + z\|} dz \leq \int_{U(0, \mathbf{y})} c_e e^{-(r - \|z\|)} dz.$$

The change of variables $r \rightarrow \rho + R_n$ yields

$$p((\rho + R_n)e_1, (\rho + R_n)e_1 + \mathbf{y}) \leq e^{-R_n} e^{-\rho} \int_{U(0, \mathbf{y})} c_e e^{\|z\|} dz.$$

If $ne^{-R_n} \rightarrow 0$, then

$$\lim_{n \rightarrow \infty} np((\rho + R_n)e_1, (\rho + R_n)e_1 + \mathbf{y}) = 0.$$

Thus,

$$\lim_{n \rightarrow \infty} e^{-np(R_n \rho e_1, R_n \rho e_1 + \mathbf{y})} = 1,$$

and therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E}\{\widehat{S}_{k,n}\} \\ = \lim_{n \rightarrow \infty} \left(n^{k+2} R_n^{d-1} e^{-(k+2)R_n} \right)^{-1} \mathbb{E}\{S_{k,n}\} = \mu_{e,k}. \end{aligned}$$

This completes the proof. \square

LEMMA 3.8. *If $f = f_e$, and $R_n \rightarrow \infty$ then*

$$\lim_{n \rightarrow \infty} \left(n^{k+3} R_n^{d-1} e^{-(k+3)R_n} \right)^{-1} \mathbb{E}\{L_{k,n}\} = \hat{\mu}_{e,k}.$$

where $\hat{\mu}_{e,k} > 0$.

PROOF. As for the proof of Lemma 3.5, mimic now the proof of Lemma 3.7, replacing T_k with an indicator function that tests whether a sub-complex generated by $k + 3$ points is connected. \square

PROOF OF THEOREM 2.4. The proof follows the same steps as the proof of Theorem 2.2. \square

3.5. *Crackle - The Gaussian Distribution.* In this section we prove Theorem 2.6.

PROOF OF THEOREM 2.6. From Lemma 3.1 we have that

$$\mathbb{E}\{S_{0,n}\} = s_{d-1}n \int_{R_n}^{\infty} r^{d-1} f(r) dr.$$

Making the change of variables $r \rightarrow (\rho^2 + R_n^2)^{1/2}$ which implies $dr = \frac{\rho}{(\rho^2 + R_n^2)^{1/2}} d\rho$, we have

$$\begin{aligned} \mathbb{E}\{S_{0,n}\} &= s_{d-1}c_g n e^{-R_n^2/2} \int_0^{\infty} (\rho^2 + R_n^2)^{(d-2)/2} \rho e^{-\rho^2/2} d\rho \\ &= s_{d-1}c_g n e^{-R_n^2/2} R_n^{d-2} \int_0^{\infty} \left((\rho/R_n)^2 + 1 \right)^{(d-2)/2} \rho e^{-\rho^2/2} d\rho. \end{aligned}$$

The integrand is bounded, and applying dominated convergence we have

$$\lim_{n \rightarrow \infty} \left(n e^{-R_n^2/2} R_n^{d-2} \right)^{-1} \mathbb{E}\{S_{0,n}\} = s_{d-1}c_g.$$

Taking $R_n = R_{0,n}^\epsilon \triangleq \sqrt{2 \log n + (d-2+\epsilon) \log \log n}$, we have

$$e^{-R_n^2/2} = n^{-1} (\log n)^{-(d-2+\epsilon)/2}$$

and so

$$\lim_{n \rightarrow \infty} n e^{-R_n^2/2} R_n^{d-2} = 0$$

which implies that

$$\mathbb{E}\{S_{0,n}\} \rightarrow 0.$$

Finally, for every $0 \leq k \leq d-1$,

$$\beta_{k,n} \leq S_{0,n}.$$

Therefore,

$$\lim_{n \rightarrow \infty} \mathbb{E}\{\beta_{k,n}\} = 0,$$

completing the proof. □

REFERENCES

- [1] Robert J. Adler, Omer Bobrowski, Matthew S. Borman, Eliran Subag, and Shmuel Weinberger. Persistent homology for random fields and complexes. *Institute of Mathematical Statistics Collections*, 6:124–143, 2010.
- [2] Lior Aronshtam, Nathan Linial, Tomasz Luczak, and Roy Meshulam. Vanishing of the top homology of a random complex. *Arxiv preprint arXiv:1010.1400*, 2010.
- [3] Eric Babson, Christopher Hoffman, and Matthew Kahle. The fundamental group of random 2-complexes. *J. Amer. Math. Soc.*, 24(1):128, 2011.
- [4] Omer Bobrowski and Robert J. Adler. Distance functions, critical points, and topology for some random complexes. *arXiv:1107.4775*, July 2011.
- [5] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fund. Math.*, 35(217-234):5, 1948.
- [6] Gunnar Carlsson. Topology and data. *American Mathematical Society. Bulletin. New Series*, 46(2):255–308, 2009.
- [7] Daniel C. Cohen, Michael Farber, and Thomas Kappeler. The homotopical dimension of random 2-complexes. *Arxiv preprint arXiv:1005.3383*, 2010.
- [8] Herbert Edelsbrunner and John Harer. Persistent homology - a survey. In *Surveys on discrete and computational geometry*, volume 453 of *Contemp. Math.*, pages 257–282. Amer. Math. Soc., Providence, RI, 2008.
- [9] Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. An introduction.
- [10] Robert Ghrist. Barcodes: the persistent topology of data. *American Mathematical Society. Bulletin. New Series*, 45(1):61–75, 2008.
- [11] Matthew Kahle. Random geometric complexes. *Discrete & Computational Geometry. An International Journal of Mathematics and Computer Science*, 45(3):553–573, 2011.
- [12] Matthew Kahle and Elizabeth Meckes. Limit theorems for Betti numbers of random simplicial complexes. *1009.4130*, September 2010.
- [13] Roy Meshulam and Nathan Wallach. Homological connectivity of random k-dimensional complexes. *Random Structures & Algorithms*, 34(3):408417, 2009.
- [14] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry. An International Journal of Mathematics and Computer Science*, 39(1-3):419–441, 2008.
- [15] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646, 2011.
- [16] Nicholas Pippenger and Kristin Schleich. Topological characteristics of random triangulated surfaces. *Random Structures & Algorithms*, 28(3):247–288, May 2006.

ROBERT J. ADLER
ELECTRICAL ENGINEERING
TECHNION, HAIFA, ISRAEL 32000
E-MAIL: robert@ee.technion.ac.il
URL: <http://web.ee.technion.ac.il/people/adler>

OMER BOBROWSKI
MATHEMATICS
DUKE UNIVERSITY, 120 SCIENCE DRIVE
DURHAM, NC 27708
E-MAIL: omer@math.duke.edu
URL: <http://www.math.duke.edu/~omer>

SHMUEL WEINBERGER
MATHEMATICS
UNIVERSITY OF CHICAGO, 5734 S. UNIVERSITY AVE
CHICAGO, IL 60637
E-MAIL: shmuel@math.uchicago.edu
URL: <http://www.math.uchicago.edu/~shmuel>